《资源科学》自由来稿的文字复制状况分析

李家永　耿艳辉　张戈丽

收稿日期: 2011-07-02
修回日期: 2012-01-11

1) 中国科学院地理科学与资源研究所《资源科学》编辑部,100101　北京市朝阳区大屯路甲 11 号,E-mail: lijy@ igsnrr. ac. cn

摘　要　　从 2009 年 1 月起《资源科学》编辑部开始使用清华同方研发的《科技期刊学术不端文献检测系统(AMLC) 》对自由来稿进行检测。结果表明,本刊来稿的文字复制率分布具有稳定性特征,复制率≤5% 的稿件占 1 /2 强,≤20% 的约占 3 /4,≥50% 的高复制率稿件多为作者自抄、重复发表或一稿多投,严格意义上的抄袭只是少数,并且不同机构投送的稿件差别显著。两年多来的应用实践说明,AMLC 是一个非常实用的检索工具,但复制与抄袭是两个不同的概念,不宜简单地用复制率评判论文的抄袭程度。

关键词　　文字复制率　科技期刊　学术不端　文献检测系统　资源科学

1　引言

近年来,我国学术界接连不断地曝出抄袭、剽窃事件丑闻。这些学术不端行为往往又和学术资源分配、学术评价、学派争斗等等问题交织在一起,使学术界的声誉受到严重损害。导致抄袭剽窃现象高频度发生的原因很多,已有不少学者发表过一些文章进行探讨,张鸣[1] 认为是学界体制问题; 肖雪慧[2] 认为最根本的是缺乏学术传统; 吴昕[3] 认为急功近利是主要病根; 胡文敏[4] 从学生视角将其归为教育体制问题; 陈桥驿[5] 更是从文化层面对学术腐败做了深入分析,认为还得从时代大背景方面探索根由,但陈先生谈的主要是国内学术独立性丧失的社会政治背景。事实上,抄袭、剽窃现象频频发生不仅仅是中国的问题,它早已成为全球性的“流行病”,美国学术信誉研究机构在 2001 年所做的调查就揭示了美国学生抄袭的严重性[6]。大量事例说明,抄袭现象泛滥是网络时代的一个产物,这与电子文档不仅传播速度快,而且复制非常便利不无关系。

基于对电子文档快捷性负面效应的认识,通过计算机系统比对检测文稿的不当引用和潜在剽窃随之展开。自 1991 年 WordCheck 软件应用以后,自然语言文本的抄袭识别技术在国外发展很快,出现了多个抄袭识别系统[7],目前国际上已经有多个文献检测系统投入运营,如 Iparadigms LLC 开发的 Turnitin (http: //www. iparadigms. com/ ,http: //www. turnitin. com/static/index. php) ,CrossRef 和 iParadigms LLC 共同研发的 Crosscheck (http: //www. crossref. org/crosscheck. html) ,Sciworth 开发的 Mydropbox(http: //www. mydropbox. com/) 等等。在国内,也有不少学者提出了多种检测算法[8-9],但目前进入实用的主要还是清华同方研发的“学术不端检测系统”[10]。

《资源科学》杂志是中国科学院地理科学与资源研究所和中国自然资源学会主办的综合性学术月刊,年发文量 300 多篇,稿源来自 3 个方面,即自由来稿、特约专稿和学会推荐稿。自 2009 年 1 月起,本刊开始使用《科技期刊学术不端文献检测系统(Academic Misconduct Literature Check, AMLC) 》对自由来稿进行检测,本文对近 3 年来的检测结果进行了初步总结和分析。

2　数据的获取与处理

本刊于 2009 年 1 月 8 日在 AMLC 创建比对数据库,选定的文献比对范围包括: 中国学术期刊网络出版总库、中国博士学位论文全文数据库、中国优秀硕士学位论文全文数据库和中国重要会议论文全文数据库等 4 个数据库; 送检稿件包括部分 2008 年的来稿,自 2009 年 1 月起开始对自由来稿进行全面检测,截至 2011 年 12 月 31 日总计送检 4997 篇。

本文使用的基础资料主要来源于从 AMLC 系统下载的检测报表(Excel 文档,包含“篇名”、“作者”、“文字复制率”、“重合字数”和“检测时间”等内容) ,并通过“稿件编号”和“作者”关联,把稿件登记表中与之匹配的“投稿日期”、“作者职称”、“单位名称”、“单位类型”等信息汇总到一起。

由于送检文稿中含有少量重检和复检稿件,也有部分特约专稿,为了减小这些特殊稿件可能带来的误差,本文对报表数据做了以下处理: (1) 删除了重检的记录; (2) 剔除了与稿件登记表不相衔接的记录; (3) 剔除了特约专稿和同文自检的记录。在以上工作基础上,考虑到统计分析的科学性和可行性,本文将时间界限严格地界定在 2009 年 1 月至 2011 年 12 月期间,进而剔除了此时段之外的所有记录。经过仔细的考查和核对,最后确定的有效记录为 4559 条。

3 检测结果的统计分析

3.1 文字复制率的分布特征

如图1所示，检测工作进行近3年来，本刊自由来稿文字复制率的分布状况并没有发生明显变化，特别是高复制率的文稿始终占有一定数量。进一步分年统计发现，2009年、2010年和2011年自由来稿的平均复制率分别为12.48%、12.86%和13.20%，中位数和众数均集中在[0，0.05]区间，标准差分别为19.93%、20.52%和19.21%，变化也不是很大。分组统计结果(图2)显示，经过一段时间的检测，文字复制率≤5%的稿件约占50%，2009、2010和2011年分别为51.17%、54.38%和47.85%，且高复制率(>50%)稿件所占的比例是呈增高趋势，复制率>50%的稿件由2009年的5.34%分别上升为7.21%和6.78%；>70%的稿件也由1.61%分别上升到3.43%和3.22%。说明检测对于遏制高复制率稿件并没有产生显著效果。

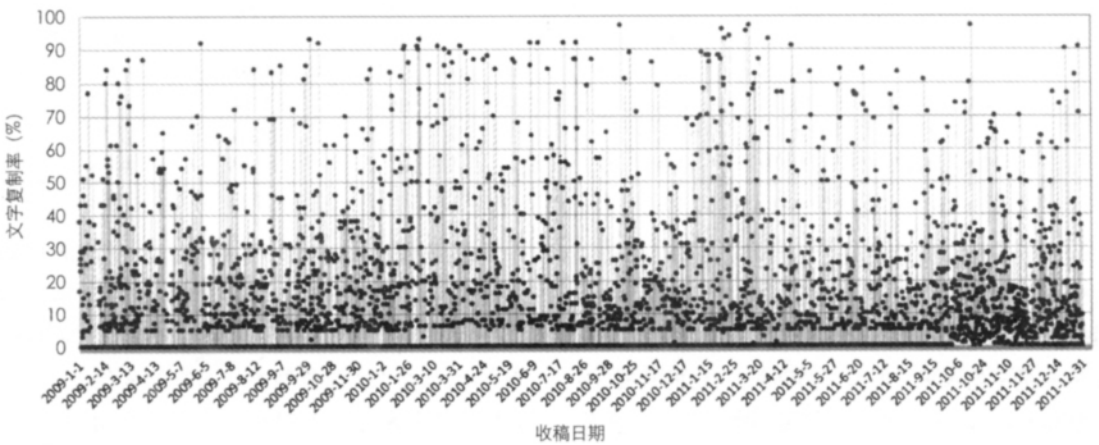

图1 2009～2011年自由来稿文字复制率的分布

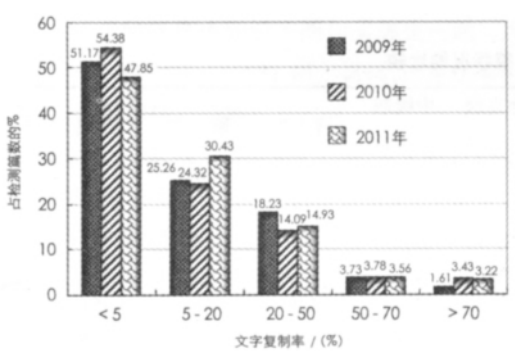

图2 分年度统计文字复制率的构成情况

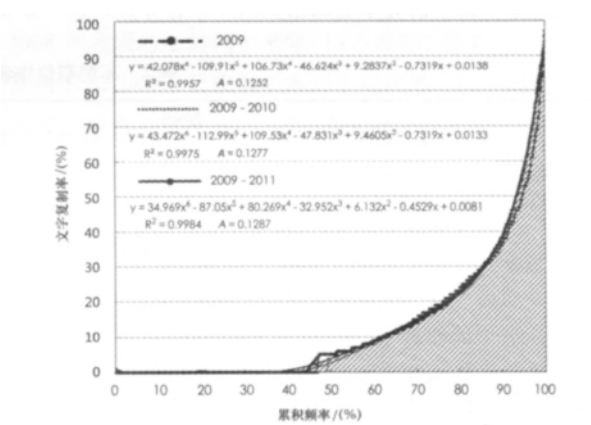

图3 文字复制率的累积分布曲线及拟合方程

累计频率曲线常常用来估计某种事物在某段时间(或区域)内出现累积效应的总体状况。下面以文字复制率为纵坐标，以小于某一复制率(即按升序排列)的累计频率为横坐标，对2009年、2009～2010年、2009～2011年等3个时期的连续系列数据进行分析。结果(图3)进一步表明，尽管随着时间的推移检测样本数量在不断地增多，但3个时期自由来稿文字复制率的累积频率曲线差别甚微，它们几乎是重合的，表现出本刊文字复制率具有稳定的分布特征：复制率≤5%的稿件始终占50%以上，<10%的占60%以上，<20%的接近80%，<50%的占90%以上，<70%的占95%以上；并且通过回归拟合方程计算得到的期望值十分接近，分别为0.1252、0.1277和0.1287。由此推测，对于稿源充裕和作者群相对稳定的期刊来说，文字复制状况主要受作者群体思维定势的影响，来稿复制率及其分布在一定时期在总体上保持相对稳定。

3.2 不同机构和作者的文字复制状况

表1是按稿件第一署名机构类型统计的结果。从表1可以看出，不同机构投送的稿件文字复制状况差别较大，在复制率>70%的133篇稿件中，没有一篇来自中国科学院，其他国

立研究机构也仅有3篇，985和211重点高校分别有21篇和25篇，其余大部分是由一般高校和地方研究机构投送的。复制率在50%～70%区间的稿件也表现出同样的规律。从不同机构来稿复制率构成比例可以更清楚地看到，中国科学院等国立研究机构来稿中复制率>20%的仅占10%左右，985和211重点高校约占20%，而一般高校则接近30%。

表1　不同机构来稿文字复制状况的比较

	文字复制率	中国科学院	其他国立研究机构	985高校	211高校	地方研究机构	一般高校	合计/平均
篇数	>70%	0	3	21	25	7	77	133
	50%～70%	3	3	29	40	4	96	175
	20%～50%	25	16	117	152	26	397	733
	<20%	293	141	774	838	134	1338	3518
	总计	321	163	941	1055	171	1908	4559
占相应篇数的%	>70%	0.00	1.84	2.23	2.37	4.09	4.04	2.92
	50%～70%	0.93	1.84	3.08	3.79	2.34	5.03	3.84
	20%～50%	7.79	9.82	12.43	14.41	15.20	20.81	16.08
	>20%累计	8.72	13.50	17.74	20.57	21.63	29.88	22.84

表2是按第一作者职位所做的统计，结果显示，研究员或副研究员身份的作者很少复制他文，在他们的来稿中极少有复制率超过50%的，复制率>20%的比例也不高，远远低于本刊22.84%的平均水平。但是，与其同等身份的教授、副教授来稿则出现另一种情况，在133篇复制率>70%的来稿中，有8篇第一署名作者是教授，14篇是副教授。同样，教授、副教授来稿中复制率>20%的稿件所占比例也高出平均水平，甚至比博士研究生的高出大约6个百分点。进一步调查发现，在国立科研机构和重点高校，往往是低职位作者复制他文的情况严重，这些单位的高复制率文稿大部分是研究生投送的；而在地方科研院所和一般高校则是高职位作者更为突出，如在133篇复制率>70%的来稿中22位教授和副教授作者多数是一般高校的。这种现象说明，一些部门和单位不切实际的定位目标以及考核和评价指标是导致复制与抄袭泛滥的一个重要因素。

表2　不同职位作者来稿文字复制状况的比较

	文字复制率	研究员	副研究员	教授	副教授	其他高级职称人员	中级职称人员	博士生	硕士生	其他	合计
篇数	>70%	0	1	8	14	4	47	17	26	16	133
	50%～70%	0	0	12	21	0	82	18	24	18	175
	20%～50%	7	4	62	93	9	228	86	146	98	733
	<20%	35	47	241	385	74	1182	519	616	419	3518
	总计	42	52	323	513	87	1539	640	812	551	4559
占相应篇数%	>70%	0.00	1.92	2.48	2.73	4.60	3.05	2.66	3.20	2.90	2.92
	50%～70%	0.00	0.00	3.72	4.09	0.00	5.33	2.81	2.96	3.27	3.84
	20%～50%	16.67	7.69	19.20	18.13	10.34	14.81	13.44	17.98	17.79	16.08
	>20%累计	16.67	9.61	25.40	24.95	14.94	23.19	18.91	24.14	23.96	22.84

3.3　高复制率稿件分析

为了更深入地认识复制现象，本文进一步利用AMLC和CNKI《中国学术文献网络出版总库》检索系统对其中复制率>50%的298篇高复制率稿件做了跟踪调查，结果见表3。

在表3列出的298篇高复制率文稿中，有95篇是作者在其会议论文、博士论文或硕士论文基础上改写的，这类稿件前期传播范围一般较小，本刊仍然作为正常来稿处理。还有32篇是作者在以往发表过的文章基础上改写的，属于重复发表，本刊一律拒收。另有40篇本刊退稿后不久(时差<3个月)即在其他杂志刊出，显然这是“一稿多投”。上述3类稿件共计167篇，约占高复制率文稿总数的60%，其共同特点是署名作者和单位基本没有变化(或者仅仅是略微调整)，但不构成占用他人成果，是否“学术不端”需要作具体的分析。除去作者自抄、自引、重复投稿和一稿多投的稿件，有抄袭嫌疑的文稿131篇，占高复制率文稿的43.95%和检测文稿总量的2.87%。事实上，就是这些高复制率的文稿还包含一些综述和评论难于定性。因此，从总体上看，在本刊自由来稿中严格意义的“抄袭论文”仍然只是少数。

表3　高复制率(>50%)稿件分类统计

	自抄自引				重复发表	一稿多投	抄袭	合计
	会议文集	博士论文	硕士论文	小计				
高复制稿篇数	19	34	42	95	32	40	131	298
各类所占%	6.38	11.41	14.09	31.88	10.74	14.42	43.96	100.00
检测后发表篇数*	8	19	17	44	23	22	40	129
其中:本刊发表	3	8	1	12	0	0	1	13
发表率(%)	42.11	55.88	40.48	46.32	71.88	55.00	30.53	43.29

* 以2011年12月统计时CNKI《中国学术文献网络出版总库》可检索到的文献为准。

需要特别指出,在这298篇高复制率的文稿中,截至2011年12月已经检索出有129篇分别在87种期刊和3个论文集上发表,有的甚至重复发表在两个以上刊物上,尤其是其中认定为"一稿多投"的40篇已经有22篇在其他期刊刊出,并且大多数和投稿原文相比没有实质性差别。笔者无意推测他刊的检测情况和采用标准,但这个问题着实提醒我们面对一个现实:凡是成形的文稿迟早都会发表,只是在不同时间出现在不同的刊物上而已。

4　典型个例分析

通过以上的统计分析,我们对本刊自由来稿文字复制状况有了一个总体的认识,但任何事物都是普遍性和特殊性的统一,只有具体问题具体分析才能切实解决问题。为此,笔者分析了一些有代表性的个例,限于篇幅,这里仅举4例。

例1　课题组成员互抄

2009年3月,某地方研究机构向本刊投送了一篇有关该省雷暴气候特征分析的文稿,此稿使用近60年的资料,分析了不同季节雷暴的空间分布和伴随天气的发生概率及其周期性特征,文稿书写流畅,图、表规范,表达清楚。但在3月4日送检时发现,此稿共计12842字,其中有10350字抄自中国气象学会2008年年会论文集收录的一篇同名文章,文字复制率高达80.59%,且与原文没有引证关系,被系统定性为"整体抄袭"。深入调查发现,原文署名作者有5位(为表达方便,本文使用字母A、B、C、D、E表示5位作者及其排序,下同),但投稿署名作者改成了3名,并新增作者F,排名顺序也改为C、A、F。按理说,像这种会议文章修改后在期刊上再次发表是有先例的。但是,由于这里不仅存在着重复发表的问题,而且存在着成果署名问题,对于期刊编辑部来说,根本无力澄清这些问题,因此不得不做退稿处理。后来,作者对文稿做了一些修改,很快就在别的期刊上发表了。

例2　自抄和重复发表

某高校教师在2010年5-8月期间8次向本刊投稿,所投7篇文稿的复制率分别为86%、61%、48%、44%、57%、56%和46%,从而引起注意。经上网检索,发现这是一位高产作者,在2008~2010年期间,他先后在42种期刊上发表了47篇文章,仅2010年就有28篇。与例1形成鲜明对照,该作者来稿署名全是独自一人,在送检的7篇文稿中,虽然篇篇都存在着严重的多源抄袭,摘抄来源文献97篇次,但复制的都是自己发表过的文章,没有一篇抄袭他人作品。像这种反复自抄重复发表,可以大大地提高作者及其单位在计量评价中的分值,也影响到基于文献计量的期刊评价。

例3　修改、伪造数据

2009年6月,某高校教师投来一篇关于资源价值估算的文稿,从表面上看,此稿有大量的"调查数据",而且论证充分。检测查出,此稿总计8743字,其中有1079字和某刊2008年发表的一篇同类文章(以下简称"前文")重合,文字复制率12.34%,只相当于本刊平均水平。但仔细比对两文发现:(1)作者部分重合;(2)署名单位没变;(3)研究方法相同;(4)研究区域来稿在甲地,前文在乙地;(5)重合部分主要出现在数字密集段落。校验这些数据察觉:两文数据时而相同时而又不同,作者在甲地收回的有效问卷为545份,乙地475份,但结果分析中竟然有20多组数据丝毫不差。很显然,在区域和样本完全不同的情况下调查结果不可能出现这么多的数据偶合,这两篇文章必有一篇数据是假的,或者都是伪造的。像这种剽窃、伪造数据的行为看起来就是几个数字,但比起前述单纯的文字复制性质要恶劣得多。

例4　检测系统问题

2009年7月,某高校教师投来一篇关于西部生态建设方面的文章,检测发现,此稿总字数11803,重合字数7435,文字复制率63%。编辑部在一周后发出退稿通知,并把检测结果以附件形式转给作者,作者随后来信申诉,信件部分内容摘抄如下:

尊敬的编辑:

您好。感谢您百忙之中就稿件问题予以答复。看了附

件的资料，情况基本清楚。拙文是基于我与另一研究者在《西部蓝皮书：中国西部经济发展报告 2005》（社会科学文献出版社出版）的基础上修改的。该书 05 年出版。去年年底因课题需要，我对文章做了些修正，但因为个人疏懒，一直拖至近期才向贵刊投稿。看了您提供的几位作者的论文，的确惊人"相似"，作为原创者，我对文章内容的来源还是自信的。这几人的文章发表时间显然晚于西部蓝皮书，我不敢妄称他人抄袭这本书，但这个事情的确令人哭笑不得。……。

类似的例子还有不少，如某研究所一名高级工程师来信，诉说他投给本刊的一篇稿件被指抄了某研究生的学位论文，而该研究生的学位论文正是在他指导下完成的，所谓的抄袭内容最先来自他的一篇会议文章。的确，目前尚有大量的图书、会议文章、学位论文、研究报告、项目申请等等文献未能收录到相关的检索系统数据库中，将来也不可能完全收录。像这种因为系统数据库不够完备，文献经无数次转抄、摘录和改写以后，检测结果完全有可能是不准确的信息，甚至混淆是非，从而增加了辨析真伪的难度。

5 讨论与结论

（1）通过对《资源科学》2009 年 1 月－2011 年 12 月自由来稿的分析，发现文字复制率具有稳定的分布特征，检测进行近 3 年来复制率 <5% 的稿件总是占 1/2 以上，<20% 的约占 3/4；<50% 的占 90% 以上，<70% 的达 95% 以上，严格意义上的抄袭论文只是少数，并且复制现象发生的频率和程度与作者群体构成密切相关。

（2）最近两年已有多篇关于 AMLC 系统应用情况的文章[11－14]发表，许多同仁强调使用该系统时要发挥编辑的主观能动性，笔者非常赞同这一观点，但不赞成制订所谓的"抄袭判别标准"。笔者认为，复制现象必定是因刊而异，因文而异。就期刊来说，作者群和稿源是关键因素，某些期刊退稿的文章照样在其他期刊发表就是明证。就文章类型而言，综述、评论必须大量旁征博引，研究报告和发布数据的文章复制率必然很低，遗憾的是因资料缺乏，本文未能完成不同文章类型复制状况的统计，有待今后继续努力。

（3）AMLC 系统是一个非常实用的检索工具，它所提供的检测结果可为查重、查新和界定引证与抄袭提供参考，但复制或转意复制的情况非常复杂，加上数据库不够完备等原因系统也有可能提供虚假信息，因而不宜简单地用复制率高低作为论文抄袭定性的依据。

（4）文字复制泛滥是一个与当代科技发展密切相关的事件，在政治、法律、道德环境建设跟不上计算机与网络技术发展速度的现实情况下，首先用技术手段解决一些问题也是合乎逻辑的。但要从根本上杜绝抄袭等学术不端行为，提高作者群体自身学术修养是关键因素。为此建议，系统研制单位和管理部门应从推进学术道德建设的高度改进系统功能，不仅用它来检测"学术不端行为"，更要为广大作者提供防患于未然的"查新"、"查重"服务，进而帮助科技工作者寻找和认准创新点，减少不必要的重复研究和重复发表。

致谢 感谢同方知网技术有限公司汪新红副总经理和王鹏副社长在 AMLC 开放前夕邀请笔者观赏系统演示，正是这次小型演示会促使笔者在使用中注意收集资料，因而才有机会获得这些体会和经验。

参考文献

1 张鸣. 学界抄袭日常化之后. 廉政瞭望，2010，(3)：68
2 肖雪慧. 学界抄袭、剽窃现象及其探源. 东方文化，2003，(6)：50－57
3 吴昕. 学术期刊中抄袭剽窃的现状分析及治理举措. 中州学刊，2008，(5)：289－291
4 胡文敏. 学生视野下的论文抄袭现象剖析. 石油教育，2009，(2)：91－93
5 陈桥驿. 论学术腐败. 学术界，2004，(5)：132－141
6 史彦军，滕弘飞，金博. 抄袭论文识别研究与进展. 大连理工大学学报，2005，45(1)：50－57
7 易彤，徐升华，万常选等. 抄袭剽窃论文识别研究综述. 情报学报，2007，28(04)：567－573
8 金博，史彦军，滕弘飞. 基于篇章结构相似度的复制检测算法. 大连理工大学学报，2007，47(01)：125－130
9 赵俊杰，胡学钢. 一种基于段落词频统计的论文抄袭判定算法. 计算机技术与发展，2009，19(4)：231－238
10 AMLC 管理办公室.《科技期刊学术不端文献检测系统(AMLC)》. http://check.cnki.net/amlc2/
11 孔琪颖，蔡斐，张利平等. 正确看待"科技期刊学术不端文献检测系统"检测结果. 编辑学报，2009，21(6)：544－546
12 史成娣. 论"学术不端文献检测系统"在编辑工作中的应用. 南昌教育学院学报，2009，24(4)：82－84
13 谭华，崔洁. 学术不端文献检测系统的使用建议. 编辑学报，2010，22(2)：153－155
14 吉家友. 学术不端文献检测系统数据分析. 中国出版，2010，31(2)：27－29